The Korean Communications
in Statistics Vol. 10 No. 3, 2003
pp. 1025-1036

# A Penalized Principal Component Analysis using Simulated Annealing

Chongsun Park[1] and Jong Hoon Moon[2]

## Abstract

Variable selection algorithm for principal component analysis using penalty function is proposed. We use the fact that usual principal component problem can be expressed as a maximization problem with appropriate constraints and we will add penalty function to this maximization problem. Simulated annealing algorithm is used in searching for optimal solutions with penalty functions. Comparisons between several well-known penalty functions through simulation reveals that the HARD penalty function should be suggested as the best one in several aspects. Illustrations with real and simulated examples are provided.

*Keywords* : Principal Component Analysis, Penalty Function, Simulated Annealing

## 1. Introduction

Dimension reduction has been an important topic in statistics and related fields for a long time and it has been very useful especially when the data sets include relatively large number of variables or features. Principal component analysis (PCA; Jolliffe, 2002) is clearly one of most frequently used method in this area and often giving relatively small number of linear combinations of variables which can effectively explain the large portion of a given data set. However, each component still include all non-zero coefficients on all variables and having problem in interpretation of the linear combination especially when the number of variables is large.

A number of methods are available to aid interpretation. A common approach is ignoring any coefficients less than some threshold value, so that the function becomes simple and the interpretation becomes easier. Jolliffe (1972, 1973) examines some of possible methods which discard irrelevant variables using multiple correlation, PCA itself, and clustering. Cadima and Jolliffe (1995) noted that this can be misleading. More formal ways of making some of the coefficients zero are to restrict the coefficients to a smaller number of possible values in the derivation of the linear functions like {-1, 0, 1} (Hausman, 1982) and variation (Vines, 2000) on this theme is also possible. Rotation method used in factor analysis is also applicable but has its drawbacks (Jolliffe, 1989, 1995). McCabe (1984) introduced a new strategy to select a

[1] Associate Professor, Department of Statistics, Sungkyunkwan University, Seoul, 110-745, KOREA
E-mail : cspark@skku.edu

[2] Graduate Student, Department of Statistics, Sungkyunkwan University, Seoul, 110-745, KOREA

subset of the variables themselves and called it 'principal variables.'

Other possible way would be introducing penalty function as in regression analysis. Recently, Jolliffe and Uddin (2001) applied $L_1$ penalty function method to maximization problem of PCA in order to force any irrelevant coefficients in the principal components to zeroes. He included $L_1$ penalty function as an extra constraint to maximization problem of variance of linear combination of variables and showed that it is more preferable to rotation methods and several others. We have seen that using $L_1$ penalty function could result in relatively severe bias for the coefficient estimates and found that hard thresholding penalty function (Antoniadis; 1997, Fan; 1997) is better in preserving original directions after adding penalty function in the model.

We compare several well-known penalty functions through simulations and real data sets and provide promising evidences that the HARD penalty function would be the best in preserving original directions of component and in other several aspects. Also we have strong feeling that the proposed method can be successfully applied to high-dimensional PCA problems with relatively large portion of irrelevant variables in selecting relevant variables.

In Section 2, basics of PCA will be introduced. Adding penalty function to PCA and modified maximization problem are in Section 3. Simulated annealing algorithm to solve above problem is included in Section 4. In Section 5, simulation studies and numerical illustrations with real data set are given. Some discussion is given in Section 6.

## 2. Principal Component Analysis

Principal component analysis (PCA; Jolliffe, 2002) is a well-known technique for dimension reduction for multivariate data sets. Several examples of its many applications include data reduction, pattern recognition, exploratory data analysis and time series prediction.

Suppose that we have $p$-dimensional data vectors $\boldsymbol{x}_j$, $n=1,\ldots,N$ and sample covariance matrix $S$ of $\boldsymbol{x}$ with $N$ observations. Usual PCA becomes solving the eigenvalue problem

$$S\boldsymbol{w}_j=\lambda_j\boldsymbol{w}_j \text{ for } j=1,\ldots,q.$$

Above problem is equivalent to the following problem so as finding unknown $p$ parameter vectors $\boldsymbol{w}_j$ which solves

$$\max_{\boldsymbol{w}_j} \boldsymbol{w}_j^T S\boldsymbol{w}_j \text{ subject to } \boldsymbol{w}_j^T\boldsymbol{w}_j=1 \text{ and } \boldsymbol{w}_h^T\boldsymbol{w}_j=0, \quad h<k.$$

Then the $q(\leq p)$ principal components of the observed vector $\boldsymbol{x}_n$ are

$$\boldsymbol{c}_j = \boldsymbol{W}^T(\boldsymbol{x}_n - \overline{\boldsymbol{x}}) \quad \text{with} \quad \boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_q)$$

such that $q$ principal axes $\boldsymbol{w}_j$ are those orthonormal axes onto which the retained variance under projection is maximal. The components $\boldsymbol{c}_n$ are then uncorrelated such that the covariance matrix $\Sigma_n \boldsymbol{c}_n \boldsymbol{c}_n^T N$ is diagonal with elements $\lambda_j$.

## 3. Penalized Principal Component Analysis

We can consider problem of extending penalized likelihood idea to the PCA for variable selection in each component. Suppose we have a penalty function $p_\lambda(\theta)$. Then the typical problems with penalty function becomes to find parameters which maximizes the following unified "Gain-Penalty" function

$$Gain(\boldsymbol{W}) - N \sum_{i=1}^{p} \sum_{j=1}^{q} p_\lambda(|w_{ij}|)$$

The first term in the above objective function may be regarded as a gain function of $\boldsymbol{W}$ for the PCA problem with $w_{ij}$ as the element of $\boldsymbol{W}$ in its $i$th row and $j$ column and $p_\lambda(\cdot)$ as a penalty function. Then PCA problem with penalty function becomes

$$\max_{\boldsymbol{w}_j} \ \boldsymbol{w}_j{}^T S \boldsymbol{w}_j - N \sum_{i=1}^{p} \sum_{j=1}^{q} p_\lambda(|w_{ij}|) \quad \text{subject to} \quad \boldsymbol{w}_j^T \boldsymbol{w}_j = 1 \text{ and } \boldsymbol{w}_h^T \boldsymbol{w}_j = 0,$$
$$h < k.$$

Fan and Li (2001) argued that unibiasedness, sparsity, and continuity as three properties that a good penalty function should have, and suggested Smoothly Clipped Absolute Deviation (SCAD) penalty function as the best one for regression problems. Several well-known penalty functions including SCAD penalty function are as follows.

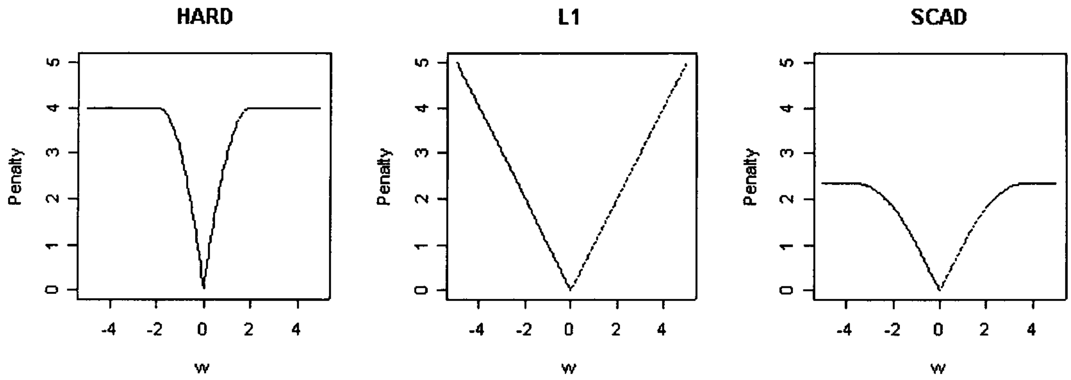
■ $L_p$: $p_\lambda(|w_{ij}|) = \lambda |w_{ij}|^p$ and it becomes LASSO ($L_1$) with $p = 1$ for least squares case.

■Hard Thresholding (HARD) Penalty: $p_\lambda(|w_{ij}|) = \lambda^2 - (|w_{ij}| - \lambda)^2 I(|w_{ij}| < \lambda)$

■Smoothly Clipped Absolute Deviation (SCAD) Penalty:

$$p_\lambda(w_{ij}) = \begin{cases} \lambda w_{ij} & \text{if} \quad w_{ij} < \lambda \\ -\dfrac{w_{ij}^2 - 2a w_{ij} + \lambda^2}{2(a-1)} & \text{if} \quad \lambda \le w_{ij} < a\lambda \\ \dfrac{(a+1)\lambda^2}{2} & \text{if} \quad w_{ij} \ge a\lambda \end{cases}$$

Unfortunately, none of three penalty functions satisfy above all three properties simultaneously.



<Figure 1> Well-known penalty functions

$L_p$ penalty function is biased and this cause some serious problem especially when applied to PCA problems. The hard thresholding (HARD) penalty function is unbiased and has sparsity but it is not continuous. SCAD behaves like something between $L_1$ and HARD and need two dimensional GCV (Generalized Cross-Validation) or usual CV to find optimal values for two parameters, $a$, and $\lambda$. Shapes of three penalty functions are in Figure 1. Simulated annealing algorithm to solve penalized eignevalue problem will be described in the next Section.

## 4. Simulated Annealing Algorithm

Simulated annealing (SA: Aarts, and Korst, 1989) method, introduced by Kirkpatric, Gelatt, and Vecchi (1983) is known to give near optimal solutions for problems with many local optimum. This method is applicable to combinatorial problems like salesman traveling problem and also to continuous multivariate optimization problems. The main idea of SA is sampling from a derived distribution according to the given objective or any other function optimize.

Now, let's define the distribution

$$u(\boldsymbol{w}) = C\exp\left(-\frac{1}{\gamma}D(\boldsymbol{w})\right)$$

with $D(\boldsymbol{w}) = -\boldsymbol{w}_j{}^T \boldsymbol{S}\,\boldsymbol{w}_j + N\sum_{i=1}^{p}\sum_{j=1}^{q} p_\lambda(|w_{ij}|)$. Then the algorithm for finding 1st principal component becomes as follows.

■[STEP 1]: Initialization

1. Set initial $\boldsymbol{w}$ from ordinary PCA

```
        2. Set $k=0$ (Step function)
        3. Set intial temparature $\gamma_0$ (Temparature at $k=0$)
        4. Set inital number of iteration $L_0$ (Number of sample at $k=0$)
        5. Set $c_0=\gamma_0$
■[STEP 2]: Repeat until convergence.
        1. For $l=1$ to $L_k$
                (1) Set $\boldsymbol{w}_{\neq w}$ from neighborhood of $\boldsymbol{w}_{old}$
                (2) Set $\boldsymbol{w}_{old}=\boldsymbol{w}_{\neq w}$ if $D(\boldsymbol{w}_{\neq w})\leq D(\boldsymbol{w}_{old})$.
                (3) Set $\boldsymbol{w}_{old}=\boldsymbol{w}_{\neq w}$ if $D(\boldsymbol{w}_{\neq w})>D(\boldsymbol{w}_{old})$
                        and $\exp([D(\boldsymbol{w}_{old})-D(\boldsymbol{w}_{\neq w})]/c_k)>U[0,1]$
        2. $k=k+1$
        3. Set $c_k=c_0\times(0.9)^k$.
■[OUTPUT]: $\boldsymbol{w}$
```

In implementing SA, it is known to be very important to set parameters and initial values for the model properly in order to get resonable solutions. For the objective function, we need to provide appropriate intial values for $\boldsymbol{w}$'s first and then it needs to choose neighborhood of them for the next iteration carefully. Also parameters in the SA algorithm should be calibrated properly, too. Here are some details for these issues.

### 4.1 Initial and iterative orthonormal $\boldsymbol{w}$'s

Initial $\boldsymbol{w}$'s: Coefficient estimates from ordinary PCA would be a very good initial values for $\boldsymbol{w}$'s in the first and each subsequent component in the penalized PCA.

Second and subsequent components should orthogonal to all previously obtained components. Hence, it should be considered in finding possible neighbor of $\boldsymbol{w}$'s for each iteration. New components should still be very close to previous one to guarantee convergence of the algorithm.

### 4.2 Parameters in penalty functions

It seems to be enough to consider $\lambda$'s which should be less than eigenvalue for each component obtained from ordinary PCA and greater than 0. It could be an option to set $\lambda$ as a function of $|g|$, absolute value of the Gain function, $p$, the number of predictors, and an appropriate multiplier, so that $\lambda=r\times|g|\sqrt{2\log(p)}$ with $r\geq 0$. Clearly, when $r=0$ it would give the same results from ordinary PCA with no constrains. And when $r \times \sqrt{2\log(p)} \ge 1$ then the algorithm forces all coefficient estimates except only one variable to zero. Hence optimal $r$ should be between 0 and $1/\sqrt{2\log(p)}$. Small experiments with simulated data sets suggest $r$ should be between 0 and 0.2 and the solution tends to include one dominating coefficient as $r$ is greater than 0.2 or so. The parameter $a$ in HARD penalty function is set at 3.7.

## 5. Illustrative Examples

We compared our method with small set of simulated and real data sets. And we considered several statistiscs to see if there would be any candidate penalty function preferable in several aspects.

### 5.1 Simulation results

For any given vector of positive real numbers and an orthogonal matrix, we can find a covariance matrix or correlation matrix whose eigenvalues are the elements of given vector, and whose eigenvectors are the columns of given matrix. The data sets are simulated based on the observation that $\boldsymbol{x}$ is marginally distributed as normal with mean $\boldsymbol{\mu}$ and covariance matrix $S = \lambda_1 \boldsymbol{w}_1 \boldsymbol{w}_1' + \lambda_2 \boldsymbol{w}_2 \boldsymbol{w}_2' + \cdots + \lambda_p \boldsymbol{w}_p \boldsymbol{w}_p'$. Further we can set $\boldsymbol{\mu}$ as zero without loss of generality.

The following sets of data are generated 100 times for each combination.

■ $N$ (number of obs.): 300, 600<br>
■ $p$ (number of var.): 5, 10<br>
■Two sets of eigenvalues and eigenvectors for each $p$.<br>
■Three sets of $r$ values for each case.<br>
■All three penalty functions for each case.

We will look at the mean and standard deviation of angles between estimated and true direction for each component. Especially, mean of zero estimates for true zero (T0: True 0) coefficients, and zero estimates for non-zero coefficients (F0: False 0) are our concern. Only parts of result are included in the two Examples following.

**Example 1.**
Eigenvalues and eigenvectors for the first example are as follows. We set three coefficients of the first two components as zero and also give relatively larger eigenvalue for the first two components. Mean and standard deviations of angles between true and estimated directions plus mean for T0 and F0 are in Table 1. Boxplots of angles for three sets of $r$ are in Figure

2.

$$
\begin{array}{lccccc}
\text{Eigenvalue:} & ( \; \mathbf{2.5}, & \mathbf{2}, & 0.3, & 0.1, & 0.1 \; ) \\
\text{Vector:} & \boldsymbol{w}_1 & \boldsymbol{w}_2 & w_3 & w_4 & w_5
\end{array}
$$

$$
\begin{bmatrix}
\mathbf{0.949} & \mathbf{0} & -0.271 & 0.11 & -0.12 \\
\mathbf{0} & \mathbf{0.949} & -0.163 & -0.184 & 0.199 \\
\mathbf{0.316} & \mathbf{0} & 0.813 & -0.331 & 0.359 \\
\mathbf{0} & \mathbf{0.316} & 0.488 & 0.551 & -0.598 \\
\mathbf{0} & \mathbf{0} & 0 & 0.735 & 0.678
\end{bmatrix}
$$

<Table 1> Eigenvalues : (2.5, 2, 0.3, 0.1, 0.1) , with N=300.

■ $r=0.08$

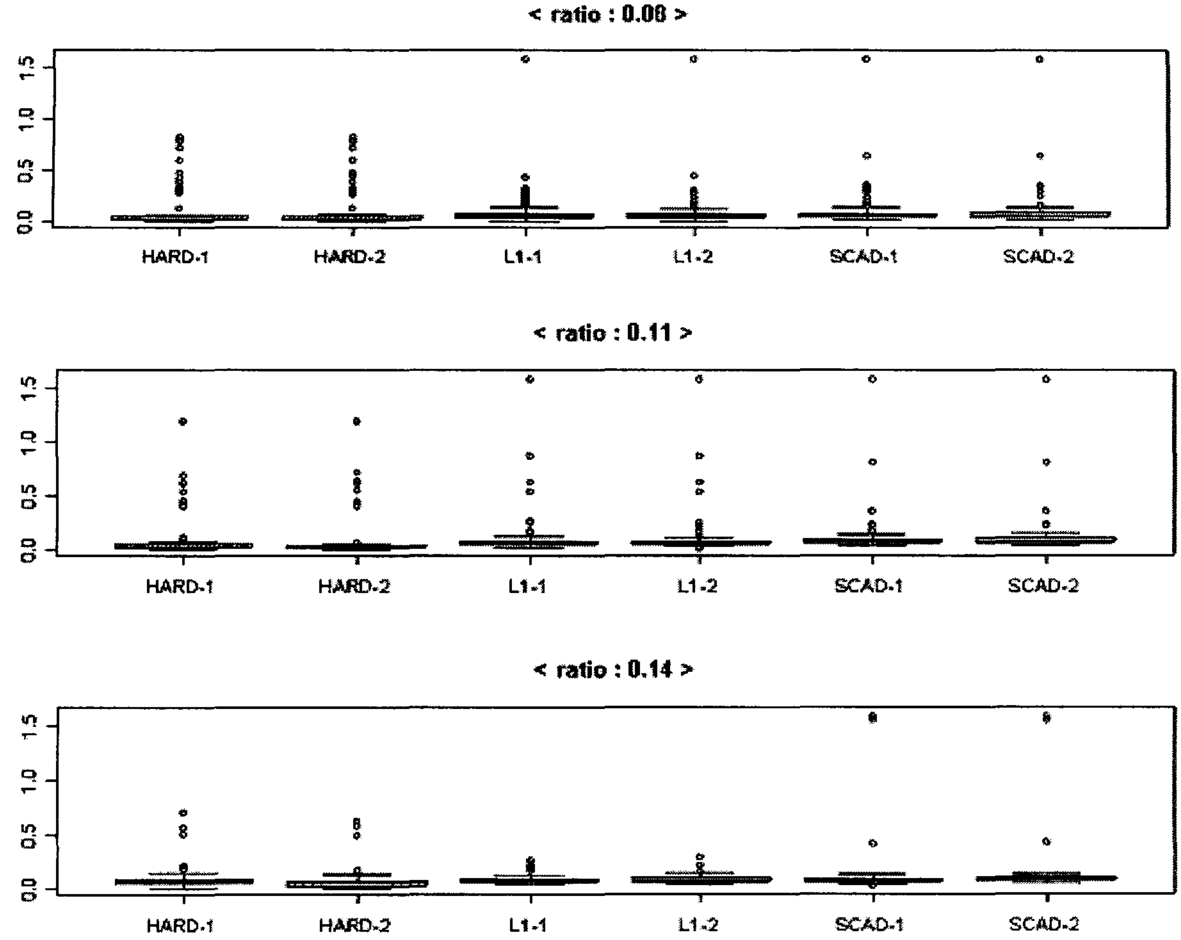
| Penal. / Comp. / Eval. | HARD | | L1 | | SCAD | |
|---|---|---|---|---|---|---|
| | 1st | 2nd | 1st | 2nd | 1st | 2nd |
| T0 | 2.65 | 2.79 | 2.48 | 2.64 | 2.57 | 2.66 |
| F0 | 0.02 | 0.02 | 0.06 | 0.06 | 0.04 | 0.04 |
| Mean | 0.097 | 0.097 | 0.111 | 0.114 | 0.103 | 0.104 |
| SD | 0.184 | 0.184 | 0.268 | 0.267 | 0.226 | 50.225 |

■ $r=0.11$

| Penal. / Comp. / Eval. | HARD | | L1 | | SCAD | |
|---|---|---|---|---|---|---|
| | 1st | 2nd | 1st | 2nd | 1st | 2nd |
| T0 | 2.8 | 2.79 | 2.65 | 2.64 | 2.69 | 2.66 |
| F0 | 0.03 | 0.03 | 0.08 | 0.07 | 0.06 | 0.06 |
| Mean | 0.099 | 0.093 | 0.140 | 0.142 | 0.129 | 0.133 |
| SD | 0.241 | 0.243 | 0.315 | 0.314 | 0.268 | 0.267 |

■ $r=0.14$

| Penal. / Comp. / Eval. | HARD | | L1 | | SCAD | |
|---|---|---|---|---|---|---|
| | 1st | 2nd | 1st | 2nd | 1st | 2nd |
| T0 | 2.91 | 2.94 | 2.86 | 2.87 | 2.8 | 2.87 |
| F0 | 0.03 | 0.02 | 0 | 0 | 0.09 | 0.09 |
| Mean | 0.083 | 0.063 | 0.074 | 0.081 | 0.151 | 0.164 |
| SD | 0.096 | 0.104 | 0.038 | 0.035 | 0.328 | 0.325 |



< ratio : 0.08 >

< ratio : 0.11 >

< ratio : 0.14 >

<Figure 2> Boxplots for angles between true and estimated directions (ratio= $r$)

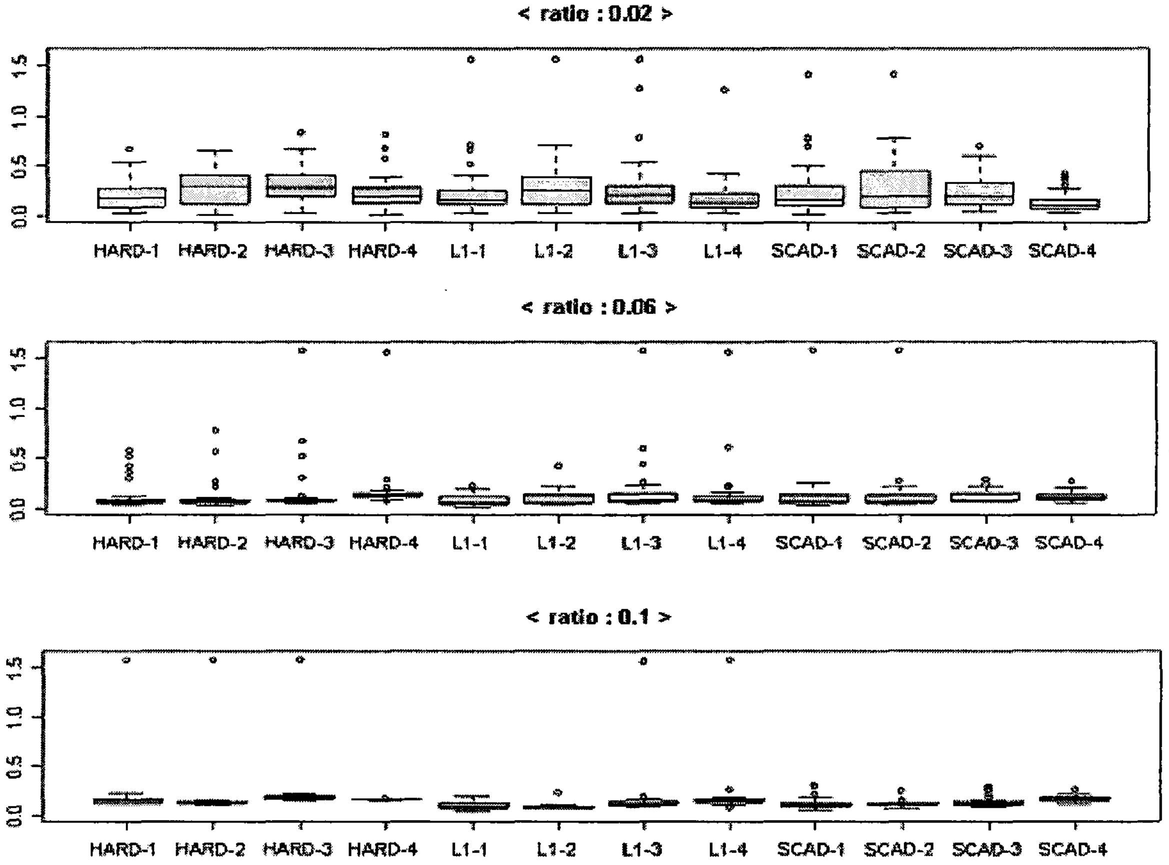
Results show that the number of true zero. increases and becomes closer to 3 as $r$ increases but F0 remains similar in all $r$ values. $L_1$ is the worst in angles between true and estimates directions for two smaller $r$'s but becomes better as it becomes larger so giving us a feeling that $L_1$ penalty seems to be robust w.r.t. $r$.

Clearly we can say that with relatively large number of observations and an appropriate $r$ our method effectively forces estimates of true zero to zero and at the same time hardly happen to give zero estimates for true non-zero coefficients. Further HARD penalty seems to be the best in preserving original directions on the wide range of the $r$ values.

**Example** 2.

Eigenvalues and eigenvectors for the second example are as follows. We set six coefficients of the first two components as zero and also give relatively larger eigenvalue for the first four components. Similar to the previous example mean and standard deviations of angles plus mean for T0 and F0 are in Table 2. Boxplots for three sets of $r$ are in Figure 3.

| Eigenvalue: | ( 3.5, | 3, | 2.5, | 2, | 0.25, | 0.25, | 0.15, | 0.15, | 0.1, | 0.1 ) |
|---|---|---|---|---|---|---|---|---|---|---|
| Vector : | $w_1$ | $w_2$ | $w_3$ | $w_4$ | $w_5$ | $w_6$ | $w_7$ | $w_8$ | $w_9$ | $w_{10}$ |
| | **0.856** | 0 | **-0.224** | 0 | 0.197 | -0.333 | -0.004 | 0 | 0.004 | 0 |
| | 0 | **0.856** | 0 | **0.231** | 0.011 | 0.191 | -0.08 | 0.218 | 0.059 | 0.222 |
| | **0.428** | 0 | **0.131** | 0 | -0.344 | 0.289 | 0.162 | 0.012 | -0.117 | 0.012 |
| | 0 | 0 | **-0.808** | 0 | -0.148 | 0.392 | -0.278 | 0.428 | 0.234 | 0.426 |
| | 0 | 0 | 0 | **0.863** | -0.181 | -0.201 | 0.2 | -0.21 | -0.234 | -0.213 |
| | 0 | **-0.428** | 0 | **0.424** | 0.085 | 0.173 | -0.473 | 0.19 | 0.467 | 0.189 |
| | **0.107** | 0 | **-0.048** | 0 | -0.04 | 0.433 | 0.453 | 0.481 | -0.467 | 0.482 |
| | **0.268** | 0 | **0.526** | 0 | -0.063 | 0.428 | -0.428 | 0.417 | 0.467 | 0.416 |
| | 0 | **-0.268** | 0 | **0.103** | 0.239 | 0.393 | 0.487 | 0.484 | -0.467 | 0.482 |
| | 0 | **0.107** | 0 | **0.11** | 0.847 | 0.142 | -0.039 | 0.221 | 0.029 | 0.222 |



<Figure 3> Boxplots for angles between true and estimated directions (ratio= $r$)

<Table 2> Eigenvalues : (3.5, 3, 2.5, 2, 0.25, 0.25, 0.15,0.15, 0.1, 0.1 ) with N=600

■ $r=0.02$

| Penal. | HARD | | | | L1 | | | | SCAD | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Eval. Comp. | 1st | 2nd | 3rd | 4th | 1st | 2nd | 3rd | 4th | 1st | 2nd | 3rd | 4th |
| T0 | 2.76 | 2.26 | 1.64 | 1.8 | 1.12 | 1.32 | 0.96 | 1.24 | 1.28 | 1.26 | 1.16 | 1.2 |
| F0 | 0 | 0 | 0.08 | 0.02 | 0.02 | 0.1 | 0.06 | 0 | 0.02 | 0.02 | 0 | 0 |
| Mean | 0.199 | 0.279 | 0.307 | 0.225 | 0.227 | 0.313 | 0.282 | 0.183 | 0.249 | 0.300 | 0.240 | 0.142 |
| SD | 0.146 | 0.174 | 0.180 | 0.167 | 0.240 | 0.311 | 0.272 | 0.182 | 0.297 | 0.297 | 0.049 | 0.041 |

■ $r=0.06$

| Penal. | HARD | | | | L1 | | | | SCAD | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Eval. Comp. | 1st | 2nd | 3rd | 4th | 1st | 2nd | 3rd | 4th | 1st | 2nd | 3rd | 4th |
| T0 | 5.36 | 5.46 | 4.78 | 4.64 | 3.12 | 3.08 | 2.76 | 2.52 | 4.24 | 4.4 | 3.82 | 3.24 |
| F0 | 0.02 | 0.02 | 1.08 | 0.68 | 0.24 | 0.38 | 0.56 | 0.08 | 0.12 | 0.16 | 0.08 | 0 |
| Mean | 0.104 | 0.099 | 0.133 | 0.161 | 0.082 | 0.095 | 0.152 | 0.139 | 0.148 | 0.152 | 0.113 | 0.119 |
| SD | 0.115 | 0.126 | 0.232 | 0.205 | 0.052 | 0.068 | 0.225 | 0.220 | 0.297 | 0.297 | 0.049 | 0.041 |

■ $r=0.1$

| Penal. | HARD | | | | L1 | | | | SCAD | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Eval. Comp. | 1st | 2nd | 3rd | 4th | 1st | 2nd | 3rd | 4th | 1st | 2nd | 3rd | 4th |
| T0 | 5.58 | 5.76 | 4.94 | 4.94 | 5.32 | 5.7 | 4.8 | 4.82 | 5.16 | 5.5 | 4.84 | 4.72 |
| F0 | 0.8 | 1.04 | 1.68 | 1.94 | 0 | 0 | 0.98 | 0.28 | 0 | 0 | 0.94 | 0.28 |
| Mean | 0.203 | 0.208 | 0.203 | 0.158 | 0.101 | 0.088 | 0.145 | 0.175 | 0.112 | 0.106 | 0.125 | 0.162 |
| SD | 0.282 | 0.346 | 0.197 | 0.003 | 0.038 | 0.023 | 0.205 | 0.202 | 0.048 | 0.024 | 0.036 | 0.023 |

When $r=0.02$ T0 is quite far from 6 but it becomes better for larger values. Behaviors of other statistics are very similar to those of previous case. $L_1$ looks best in mean angles for all values of $r$ but has relatively larger variance than HARD.

Overall, it looks reasonable to use HARD for the PCA problem since it looks best in forcing coefficients of irrelevant variables to zero and at the same time in preserving original directions after introducing penalty functions.

## 5.2 Real example

Here we will look at the results of applying penalized PCA method to well-known IRIS data set (Fisher, 1936). We combine three kinds of iris data sets into one and applied ordinary and penalized PCA. There are four variables, Sepal Length, Sepal Width, Petal Length, and Petal Width. Each species have 50 observations so total of 150 cases. We look at coefficients estimates for first two components from original and penalized PCA. We tried $r$={0.02, 0.1, 0.18} and only reported results for $r=0.1$ in the Table 3.

<Table 3> Results for IRIS data set with $r=1.0$

■ $r=0.1$

| Penal.<br>Comp.<br>Variable | PCA | | HARD | | L1 | | SCAD | |
|---|---|---|---|---|---|---|---|---|
| | 1st | 2nd | 1st | 2nd | 1st | 2nd | 1st | 2nd |
| Sepal Length | 0.3614 | 0.6566 | 0.3665 | 0.5860 | 0.3513 | 0.5212 | 0.3433 | 0.6003 |
| Sepal Width | -0.0845 | 0.7302 | 0 | 0.7709 | -0.0416 | 0.8362 | -0.0407 | 0.7745 |
| Petal Length | 0.8567 | -0.1734 | 0.8581 | -0.2498 | 0.8713 | -0.1706 | 0.8773 | -0.1992 |
| Petal Width | 0.3583 | -0.0755 | 0.3596 | 0 | 0.3402 | 0 | 0.3328 | 0 |

In the 1st component Sepal Width and Petal Width have 0 coefficient for the HARD penalty and 0 for Petal Width only with other two penalty functions. HARD penalty function seems to be the best in preserving original directions for all components. With $r=0.18$ (results not included here) some coefficients tends to dominate so that coefficient estimates related to variables with small estimates becomes smaller and resulted with 0 for Petal Length in $L_1$ and SCAD penalty functions. We get all non-zero coefficient estimates with $r=0.02$ which is very close to ordinary PCA with no penalty function.

## 6. Discussions

In this paper we propose a variable selection method for principal component analysis. We incorporated penalty functions for each coefficient estimates and solve penalized optimization problem using simulated annealing algorithm.

According to results from simulated and real data sets we found our method turned out to be very effective in forcing coefficient estimates zero for irrelevant variables in each component and further HARD penalty function seems to be preferable with relatively small bias for wide range of $r$ values than well-known SCAD and $L_1$ penalty functions. Hence, the proposed method can be successfully applied to high-dimensional PCA problems with relatively large portion of irrelevant variables in the data set.

More detailed research for theoretical results regarding properties and asmptotics for coefficient estimates and further study to find optimal values for parameters like $r$, and $a$ in the penalty function as an example by using cross-validation (CV) or generalized CV are necessary.

## References

[1] Aarts, E., and Korst, J. (1989). *Simulated Annealing and Boltzmann Machines,* John Wiley & Sons.

[2] Antoniadis, A. (1997). Wavelets in Statistics : A Review, *Journal of the Italian Statistical Association*, Vol. 6, 97-144.

[3] Cadima, J., and Jolliffe, I. T. (1995). Loadings and Correlations in the Interpretation of Principal Components, *Journal of Applied Statistics*, Vol. 22, 203-214.

[4] Fan, J., and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties, *Journal of the American Statistical Association*, Vol. 96, 1348-1360.

[5] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems, *Annals of Eugenics,* Vol. 7, 179-188.

[6] Hausman, R. (1982). *Constrained Multivariate Analysis. In Zanckis, S. H. and Rustagi, J. S. (Eds.). Optimisation in Statistics*, 137-151, North Holland: Amsterdam.

[7] Johnson, R. A., and Wichern, D. W. (2002). *Applied Multivariate Statistical Analysis,* Prentice Hall.

[8] Jolliffe, I. T. (1989). Rotation of Ill-defined Principal Components, *Applied Statistics*, Vol. 38, 139-147.

[9] Jolliffe, I. T. (1995). Rotation of Principal Components: Choice of Normalization Constraints, *Journal of Applied Statistics,* Vol. 22, 29-35.

[10] Jolliffe, I. T., and Uddin, M. (2001). A Modified Principal Components Technique Based on the LASSO, *Unpublished Manuscript.*

[11] Jolliffe, I. T. (2002). *principal component analysis*, Springer.

[12] Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by Simulated Annealing, *Science*, Vol. 220 Issue 4598.

[13] Jolliffe, I. T., and Uddin, M. (2002). A Modified Principal Component Technique Based on the LASSO, PostScript preprint, Department of Mathematical Sciences, University of Aberdeen.

[14] Jolliffe, I. T. (1972). Discarding variables in a Principal Component Analysis Ⅰ:Artificial Data, *Applied Statistics*, Vol. 21, 160-173

[15] Jolliffe, I. T. (1973). Discarding variables in a Principal Component Analysis Ⅱ:Real Data, *Applied Statistics*, Vol. 22, 21-31.

[16] McCabe, G. P. (1984). Principal Variables, *Technometrics*, Vol. 26, 137-144.

[17] Vines, S. K. (2000). Simple Principal Components, *Applied Statistics*, Vol. 49, 441-451.

[ Received July 2003, Accepted October 2003 ]